

Algorithmic data fusion methods for tuberculosis

Cagri Ozcaglar

Rensselaer Polytechnic Institute
Department of Computer Science



Ph.D. Thesis Defense
7/5/2012



Contributions

1. TCF: Tensor Clustering Framework

- A new sublineage structure of MTBC strains using multiple biomarkers
- Genomic data fusion via multiple-biomarker tensors

2. Evolution model of spoligotypes

- Evolutionary analysis of spoligotypes using multiple biomarkers
- Genomic mutation mechanism fusion

3. UBF: Unified Biclustering Framework

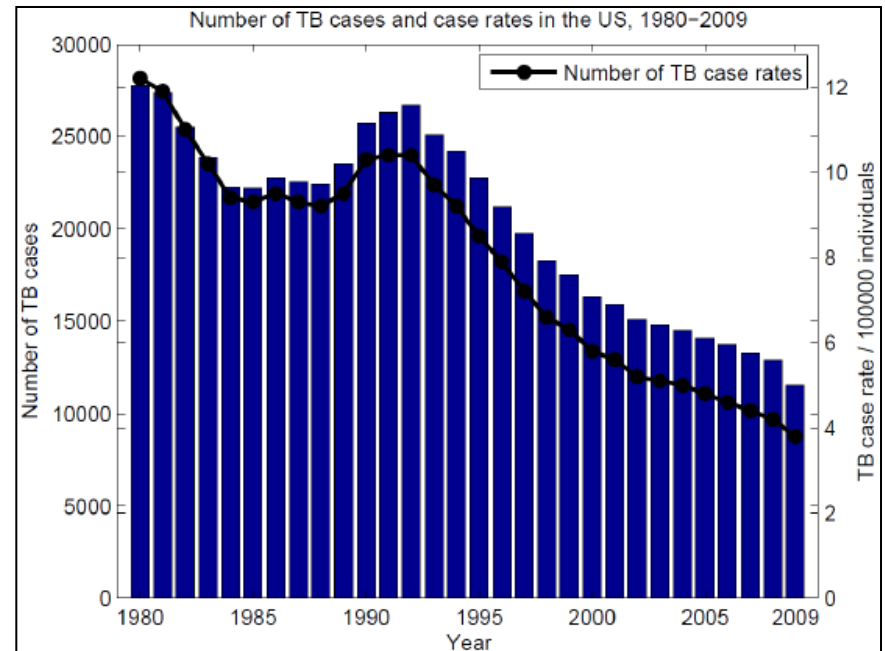
- Host-pathogen association analysis of tuberculosis patients
- Genome-phenome data fusion

Outline

1. Introduction: TB and MTBC
2. Background: Post-genomic data analysis
3. TCF: Tensor Clustering Framework
4. Evolution model for spoligotypes
5. UBF: Unified Biclustering Framework
6. Conclusion

TB: Tuberculosis

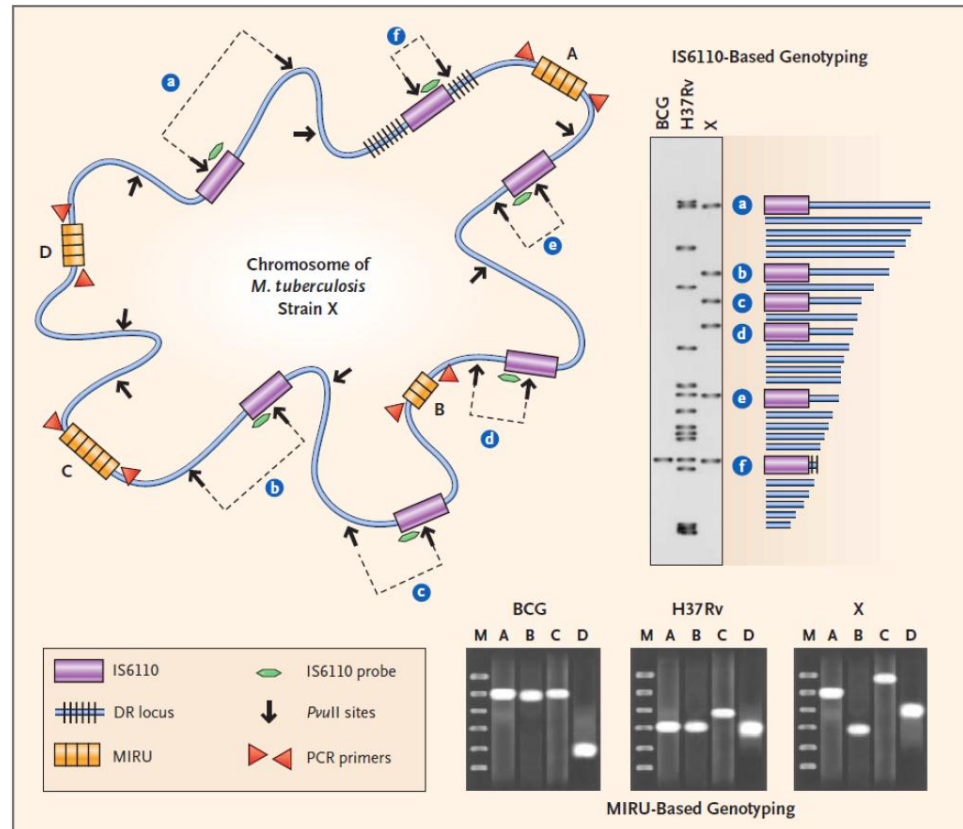
- Infectious disease
 - Airborne infection or transmission
 - 1 / 3 of the human population infected with TB
 - 90% of TB cases remain latent
-
- 1980-2009: TB cases decrease
 - Exception: Early 1990s
 - > 2 million/year die from TB



Ozcaglar et al., Epidemiological models of *Mycobacterium tuberculosis* complex infections, *Mathematical Biosciences*, 2012.

MTBC: *M. tuberculosis* complex

- MTBC bacteria: causative agent of TB
- Genotyped by multiple biomarkers:
 - Spoligotype
 - MIRU-VNTR
 - RFLP
 - SNPs
 - LSPs



Barnes et al., *New England J. Medicine*, 2003

Motivation

- Multiple sources of data from:

- MTBC strains
- TB patients

- To solve the following problems:

1. MTBC differentiation
 - Using multiple biomarkers
2. Evolutionary analysis of an MTBC biomarker
 - Using an additional biomarker
3. Host-pathogen association analysis
 - Incorporating distance and time

Genomic data fusion

Genomic data fusion

Genome-phenome data fusion

- Algorithmic data fusion methods:

1. TCF: Tensor Clustering Framework
2. SpolTopol: Spoligoforest Topology analysis
3. UBF: Unified Biclustering Framework

Outline

1. Introduction: TB and MTBC
2. Background: Post-genomic data analysis
 - Classification and Clustering
 - Biclustering
 - Multiway modeling
 - Phylogenetic analysis
3. TCF: Tensor Clustering Framework
4. Evolution model for spoligotypes
5. UBF: Unified Biclustering Framework
6. Conclusion

Classification and Clustering

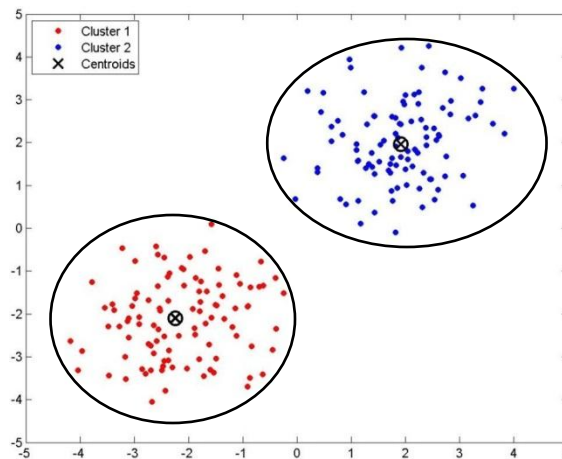
1. Classification

- Predict classes of data points
- *Supervised learning*: Classes known *a priori*

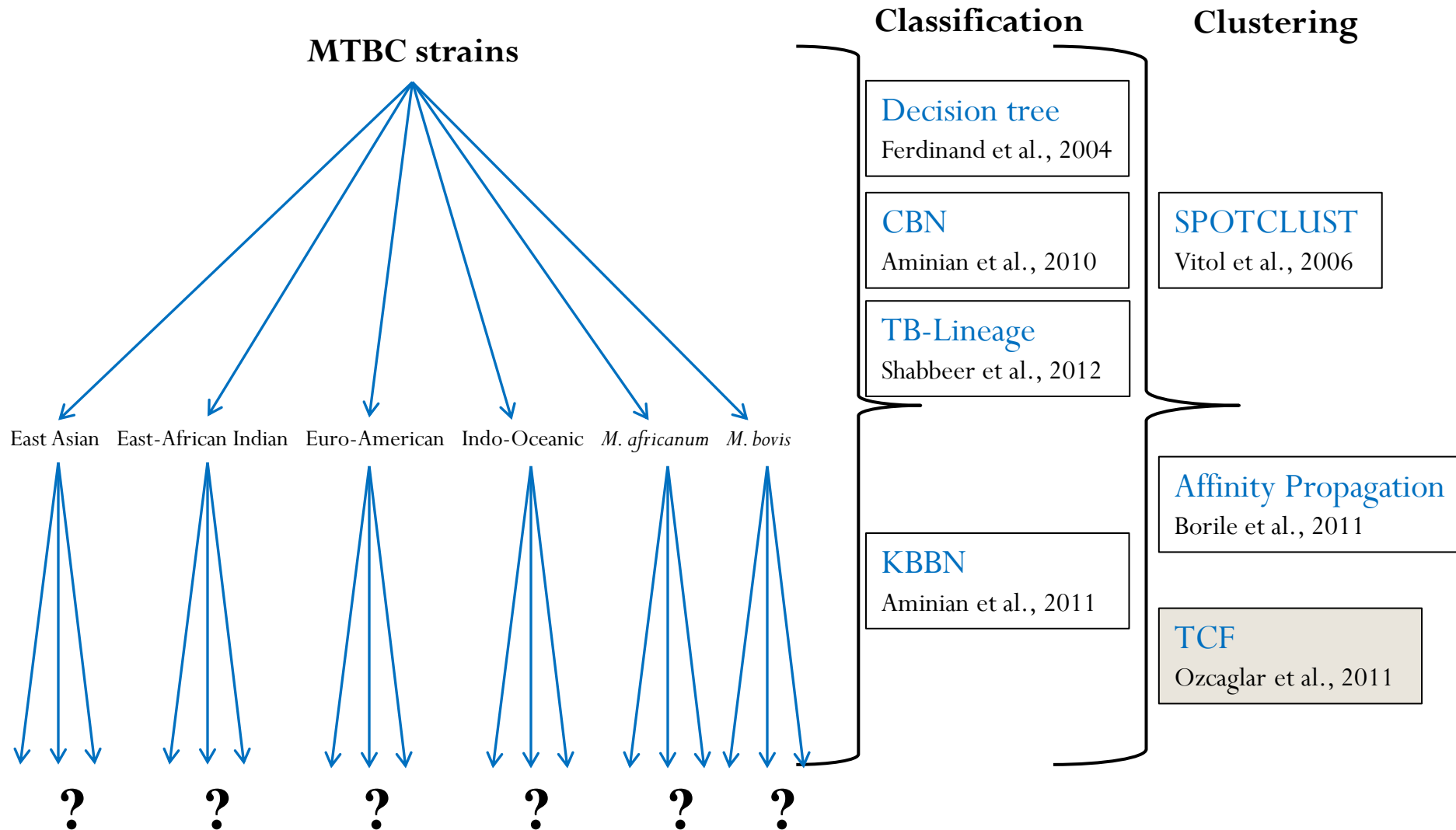


2. Clustering

- Grouping data points
- *Unsupervised learning*: Classes unknown *a priori*

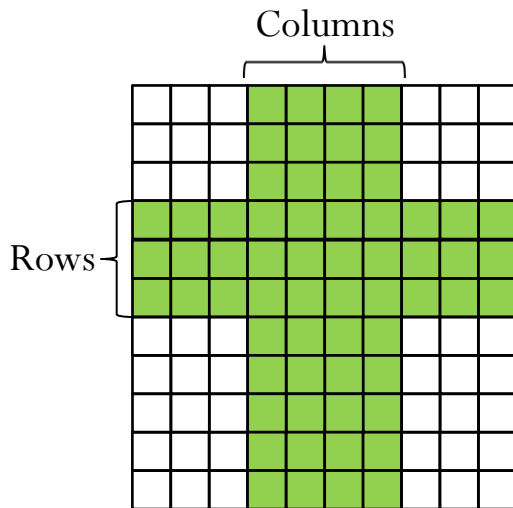


Classification and Clustering of MTBC



Biclustering

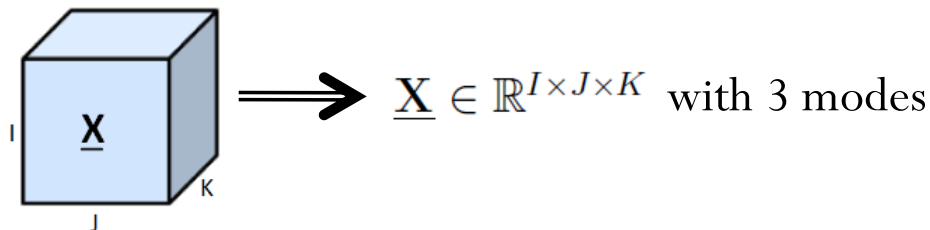
- **Biclustering**: Clustering rows and columns simultaneously
- Concept coined by Hartigan (1972)
- Term used by Mirkin (1996)
- Commonly used for microarray data analysis in 2000s
- Find a submatrix within the data matrix



- **Biclustering algorithms**:
 - Cheng and Church: Row/column add/remove
 - CTWC: Coupled Two-Way Clustering
 - SAMBA: Statistical-Algorithmic Method for Bicluster Analysis
 - BiMax: Binary Inclusion-Maximal algorithm
 - OPSM: Order-Preserving Submatrix algorithm

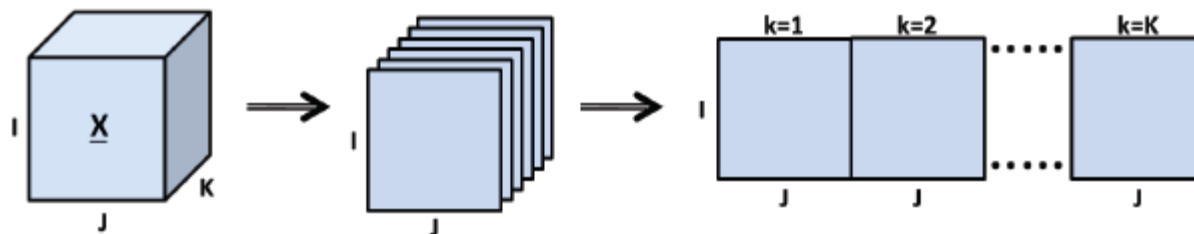
Multiway modeling: terminology

- Tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N}$ has N modes.



- **Matricization:** Unfolding

- Mode- n matricization of a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \dots \times I_N} : \mathbf{X}_{(n)}$



- **Kronecker product**

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}$$

- **Khatri-Rao product**

$$\mathbf{A} \odot \mathbf{B} = [a_1 \otimes b_1 \quad a_2 \otimes b_2 \quad \dots \quad a_K \otimes b_K]$$

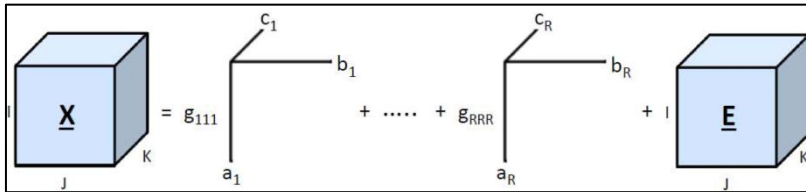
Multiway models and algorithms

Models

PARAFAC

$$\underline{\mathbf{X}} \approx \sum_{r=1}^R \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r = [\mathbf{A}, \mathbf{B}, \mathbf{C}]$$

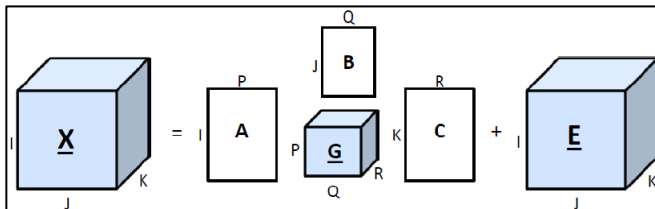
$$\mathbf{X}_{(1)} = \mathbf{A} (\mathbf{C} \odot \mathbf{B})' + \mathbf{E}_{(1)}$$



Tucker3

$$\underline{\mathbf{X}} \approx \sum_{p=1}^P \sum_{q=1}^Q \sum_{r=1}^R g_{pqr} \mathbf{a}_p \circ \mathbf{b}_q \circ \mathbf{c}_r = [\underline{\mathbf{G}}; \mathbf{A}, \mathbf{B}, \mathbf{C}]$$

$$\mathbf{X}_{(1)} = \mathbf{A} \mathbf{G}_{(1)} (\mathbf{C} \otimes \mathbf{B})' + \mathbf{E}_{(1)}$$



Algorithms

PARAFAC-ALS

Algorithm 1 PARAFAC-ALS($\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}, R$)

- 1: Initialize $\mathbf{A} \in \mathbb{R}^{I \times R}, \mathbf{B} \in \mathbb{R}^{J \times R}, \mathbf{C} \in \mathbb{R}^{K \times R}$
- 2: **while** (convergence criterion) **do**
- 3: $\mathbf{Z} = \mathbf{C} \odot \mathbf{B}$
 $\mathbf{A} = \mathbf{X}_{(1)} \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1}$
- 4: $\mathbf{Z} = \mathbf{C} \odot \mathbf{A}$
 $\mathbf{B} = \mathbf{X}_{(2)} \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1}$
- 5: $\mathbf{Z} = \mathbf{B} \odot \mathbf{A}$
 $\mathbf{C} = \mathbf{X}_{(3)} \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1}$
- 6: **end while**

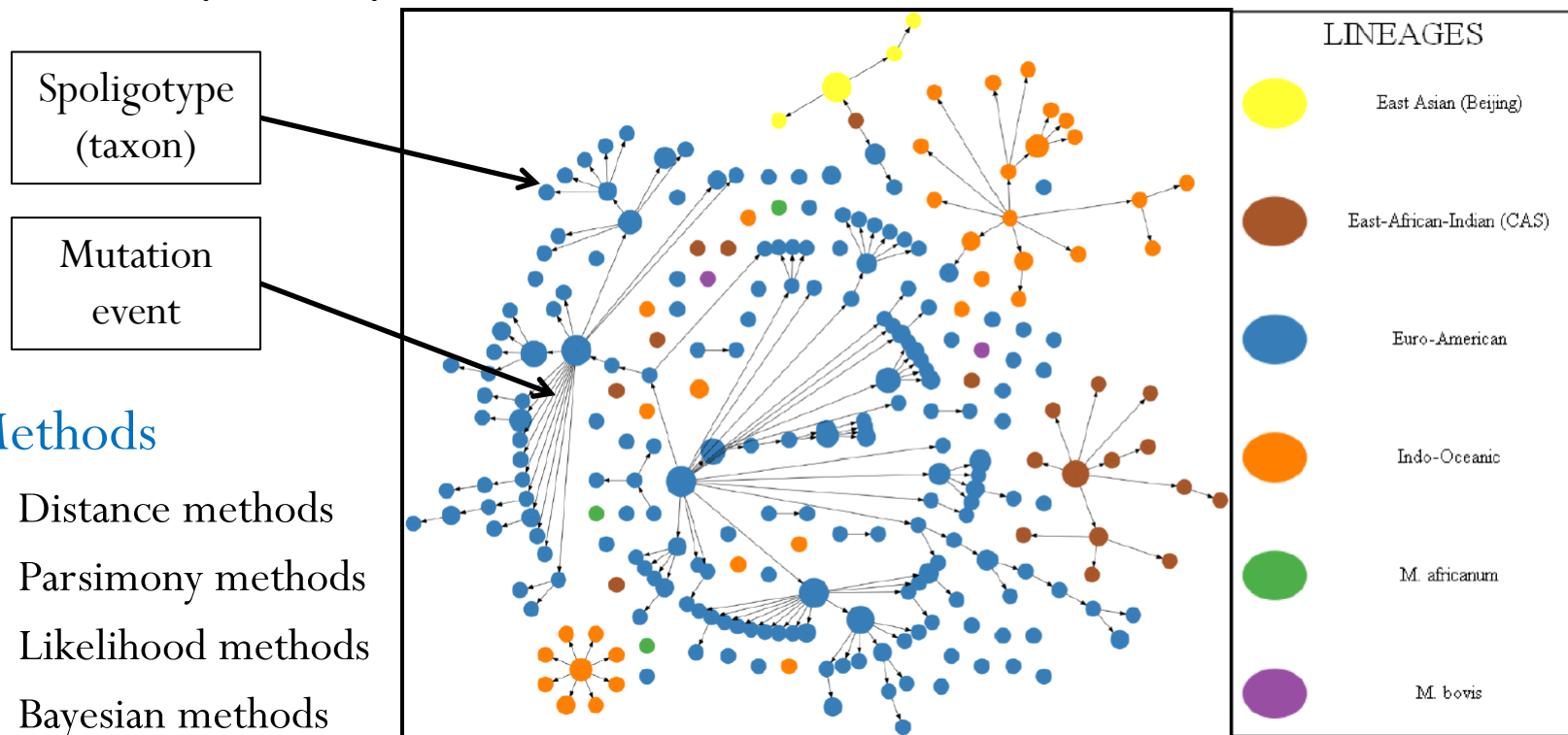
Tucker3-ALS

Algorithm 2 Tucker3-ALS($\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}, [P, Q, R]$)

- 1: Initialize $\mathbf{A} \in \mathbb{R}^{I \times P}, \mathbf{B} \in \mathbb{R}^{J \times Q}, \mathbf{C} \in \mathbb{R}^{K \times R}$
- 2: **while** (convergence criterion) **do**
- 3: $\mathbf{Z} = \mathbf{X}_{(1)} (\mathbf{C} \otimes \mathbf{B})$
 $\mathbf{A} = \text{SVD}(\mathbf{Z}, P)$
- 4: $\mathbf{Z} = \mathbf{X}_{(2)} (\mathbf{C} \otimes \mathbf{A})$
 $\mathbf{B} = \text{SVD}(\mathbf{Z}, Q)$
- 5: $\mathbf{Z} = \mathbf{X}_{(3)} (\mathbf{B} \otimes \mathbf{A})$
 $\mathbf{C} = \text{SVD}(\mathbf{Z}, R)$
- 6: **end while**
- 7: $\mathbf{G}_{(1)} = \mathbf{A}' \mathbf{X}_{(1)} (\mathbf{C} \otimes \mathbf{B})$

Phylogenetic analysis

- **Phylogeny:** Reconstruction of evolutionary history of a group of organisms, *taxa*.
- **Phylogenetic tree:** The graphical structure that represents inferred evolutionary history of *taxa*.



- **Methods**

- Distance methods
- Parsimony methods
- Likelihood methods
- Bayesian methods

Outline

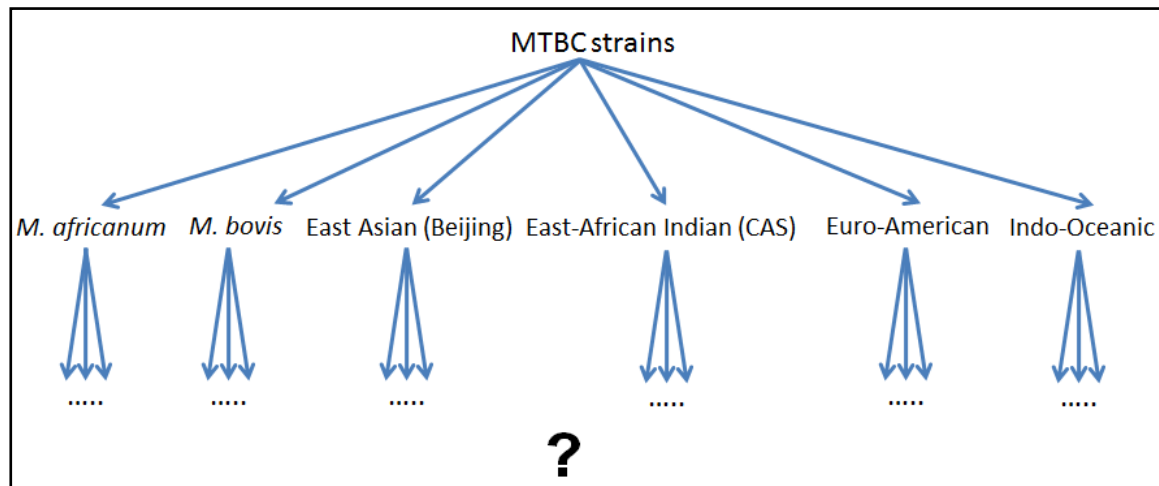
1. Introduction: TB and MTBC
2. Background: Post-genomic data analysis
3. TCF: Tensor Clustering Framework
 - [Ozcaglar et al., IEEE BIBM, 2010]
 - [Ozcaglar et al., BMC Genomics, 2011]
4. Evolution model for spoligotypes
5. UBF: Unified Biclustering Framework
6. Conclusion

Motivation: TCF

- Why do we cluster? MTBC strains vary in:
 - Infectivity
 - Host-pathogen association (e.g. Mexico, Indo-Oceanic)
 - Transmissivity (e.g. W-Beijing)
 - Virulence [Gagneux et al., *PNAS* 2006]
 - Drug resistance
- Classification of MTBC strains into major lineages:
 - Characteristics of MTBC strains
 - Unusual traits of MTBC strains
- Further **subdivide MTBC major lineages**
 - **Find more specific groups** of MTBC strains
- Use **multiple biomarkers**
 - Spoligotypes
 - MIRU patterns

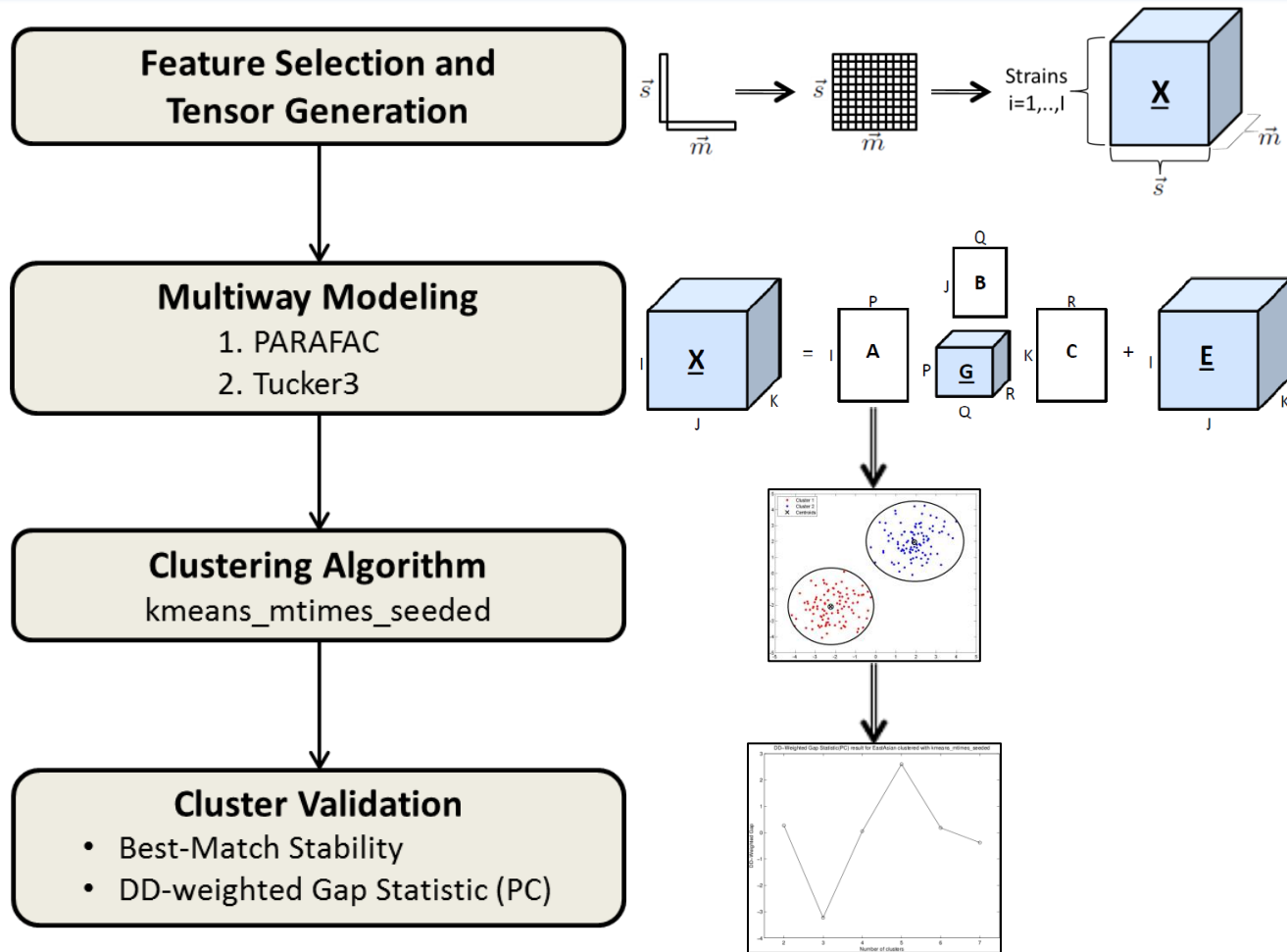
Goal: TCF

- **Goal:** Divide major lineages into sublineages
 - Using multiple biomarkers via genomic data fusion



- **Need:** A method to cluster strains
 - Using multiple biomarkers simultaneously
- **Tool:** The Tensor Clustering Framework (TCF)
 - Using Multiple-Biomarker Tensors (MBT)

TCF: Tensor Clustering Framework



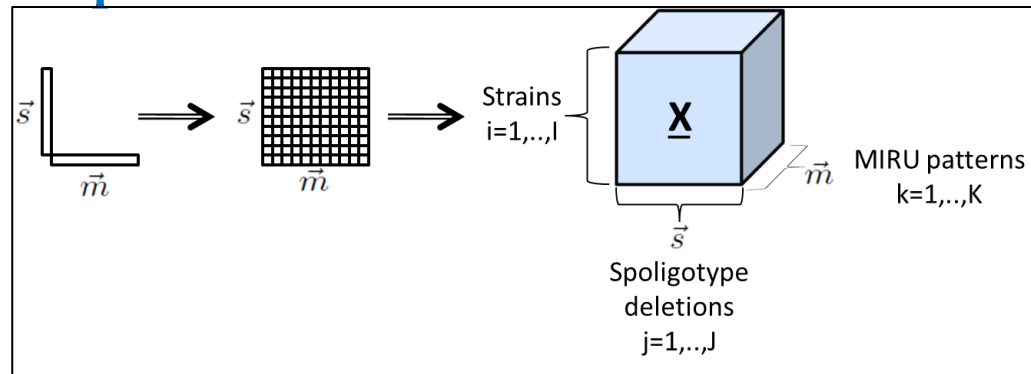
Ozcaglar et al., Sublineage structure analysis of *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors, *BMC Genomics*, 2011.

MBT: Multiple-Biomarker Tensor

- Biomarker kernel matrix

- Spoligotype deletion vector: \vec{s} , binary vector.
- MIRU pattern vector: \vec{m} , digit vector.
- Biomarker kernel matrix: $\vec{s} \times \vec{m}$, outer product of \vec{s} and \vec{m} .

- MBT: Multiple-biomarker tensor



- $\underline{X}_{ijk} = \delta_{ij} r_{ik}$: Coexistence of spoligotype deletions with MIRU loci



$$\delta_{ij} = \begin{cases} 0, & \text{if spoligotype deletion } j \text{ does not occur in strain } i, \\ 1, & \text{if spoligotype deletion } j \text{ occurs in strain } i. \end{cases}$$

r_{ik} = Number of repeats in MIRU locus k of strain i

Clustering algorithm and validation

- K-means is a commonly used clustering algorithm
- Two improvements to weaknesses:
 1. **Initial Centroids problem:** Initial centroids are chosen randomly.
 - Careful seeding using `kmeans++` [Arthur et al., *SODA*, 2007]
 2. **Local Minima problem:** The objective function can fall into local minima.
 - Repeat k-means multiple times, retrieve the run with minimum objective.

Algorithm 4 `kmeans++(A, k)`

```
1: Pick the first centroid  $c_1$  at random: InitCentroids = {c1}
2: for  $i = 2$  to  $k$  do
3:   Find  $D(a)$ , distance to the closest centroid picked so far, for each data point
    $a \in A$ 
4:   Pick the data point  $a$  with maximum  $D(a)$  as new centroid
    $c_i = \arg \max_a D(a)$ 
5:   Add  $c_i$  to the set of initial centroids:
   InitCentroids = InitCentroids  $\cup$  {ci}
6: end for
7: Run kmeans(A, k) with InitCentroids
```

Algorithm 5 `kmeans_mtimes_seeded(A, k, m)`

```
1: for  $i = 1$  to  $m$  do
2:   kmeans++(A, k)
3:   Get the objective value of k-means run  $i$ 
4: end for
5: Pick the k-means run with the minimum objective value
```

• Cluster validation

- Best-match stability
- DD-weighted gap statistic

The Dataset

- 6848 distinct MTBC strains
 - Spoligotype and 12-loci MIRU.
 - CDC + MIRUVNTR_{plus}
 - The strains are labeled by [major lineages](#) and [SpolDB4 lineages](#).

Spoligotype	MIRU	Major lineage	SpolDB4 lineage
111100000000111111111111111111111111101111	225424243522	<i>Mycobacterium africanum</i>	AFRI
0000000000000000111111111101111111111100000	235324253421	<i>Mycobacterium bovis</i>	BOV
00100011111	223325163533	East Asian	BEIJING
0000000111111111111111110000000000000111111111	225425163534	East-African Indian	CAS
11111111111111111111111100000000000000110111111	144426221234	Indo-Oceanic	EAI
0011111111111111111111111101000000100001111111	224322153322	Euro-American	H1

Major lineage	# Strains	# Spoligotype deletions
<i>M. africanum</i>	64	22
<i>M. bovis</i>	102	34
East Asian (Beijing)	571	5
East-African Indian(CAS)	508	18
Indo-Oceanic	1023	28
Euro-American	4580	109

Results: Tensor sublineages

- Apply TCF on MBT of each major lineage
- Number of components used in PARAFAC and Tucker3 on MBT

Major Lineage	Tensor size	PARAFAC		Tucker3	
		# Components	CC / Variance	# Components	Variance
<i>M. africanum</i>	64 × 22 × 12	3	95.08 / 93.33	[4 3 1]	91.94
<i>M. bovis</i>	102 × 34 × 12	2	100.00 / 86.02	[7 5 1]	91.05
East Asian (Beijing)	571 × 5 × 12	2	100.00 / 81.58	[3 4 2]	93.09
East-African Indian (CAS)	508 × 18 × 12	3	90.75 / 80.48	[6 6 4]	94.27
Indo-Oceanic	1023 × 28 × 12	5	92.99 / 80.35	[15 13 5]	95.55
Euro-American	4580 × 109 × 12	14	99.06 / 89.83	[14 13 5]	89.77

- Number of tensor sublineages and validation measure values

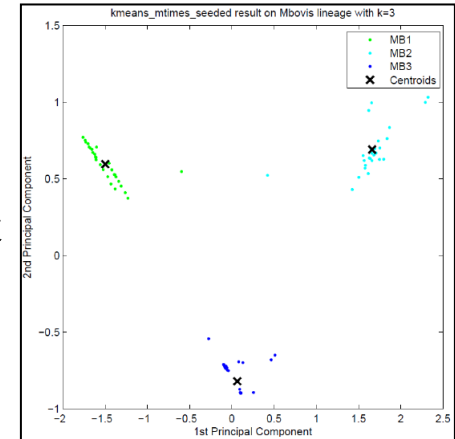
Major Lineage	# SpoIDB4 families	# Tensor sublineages	F-measure	Stability
<i>M. africanum</i>	4	4	0.66	1
<i>M. bovis</i>	5	3	0.71	1
East Asian (Beijing)	2	6	0.88	1
East-African Indian (CAS)	4	4	0.75	1
Indo-Oceanic	13	9	0.67	0.86
Euro-American	33	35	0.53	0.84

Subdivision of *M. bovis* lineage

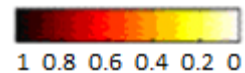
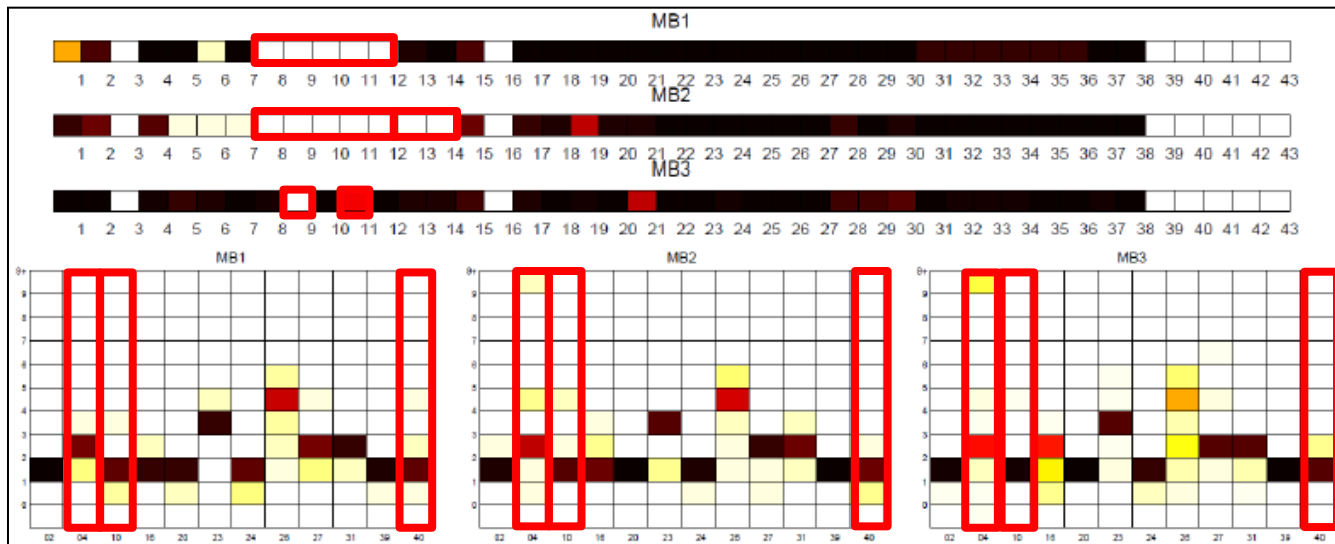
Confusion matrix

	MB1	MB2	MB3
Stability	1	1	1
BOV	7	5	5
BOVIS1	0	0	29
BOVIS1_BCG	0	0	11
BOVIS2	24	0	0
BOVIS3	0	21	0

PCA plot



Biomarker signature



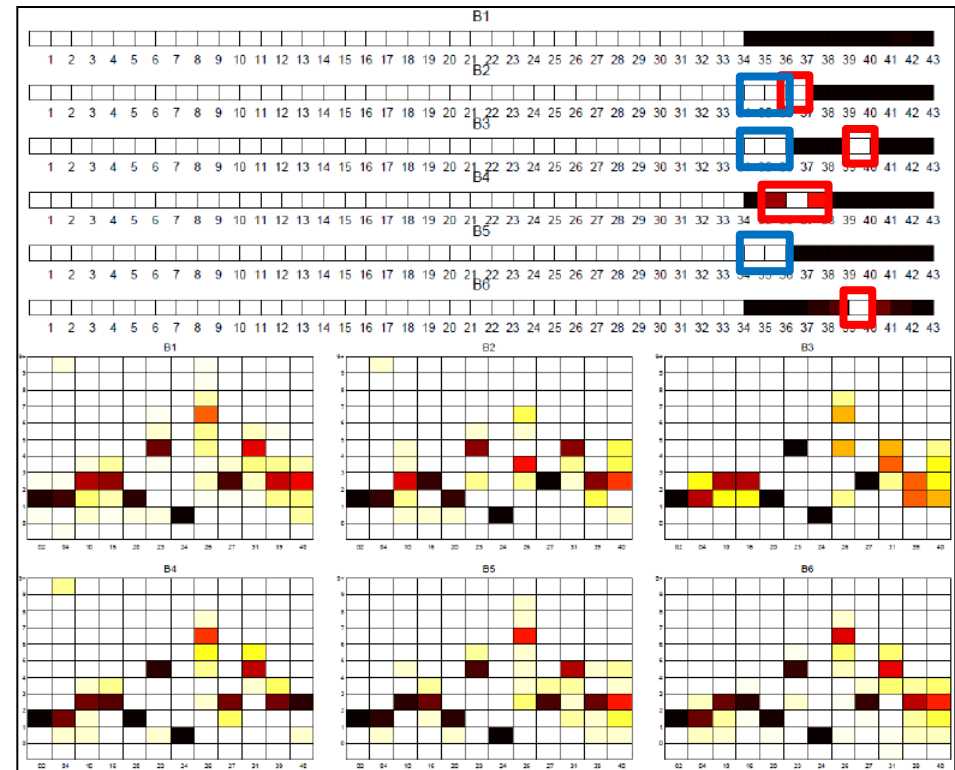
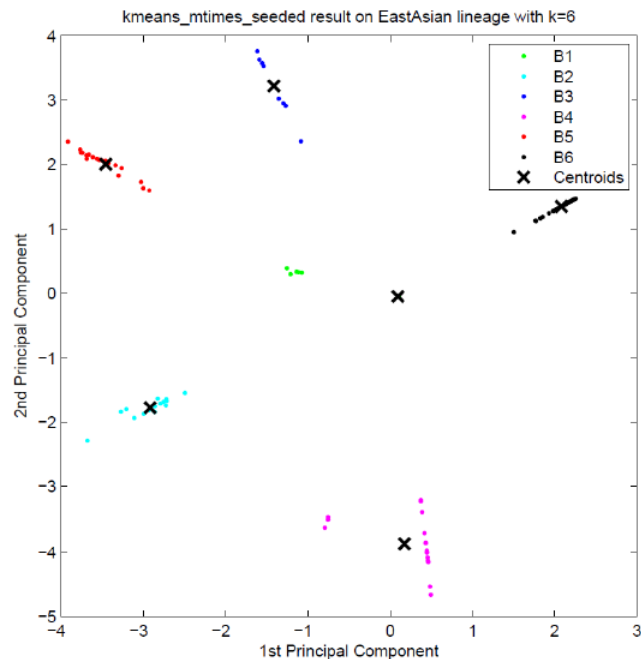
Subdivision of East Asian lineage

Confusion matrix

	B1	B2	B3	B4	B5	B6
Stability	1	1	1	1	1	1
BEIJING	468	0	0	18	0	41
BEIJING-LIKE	0	16	8	0	20	0

Biomarker signature

PCA plot

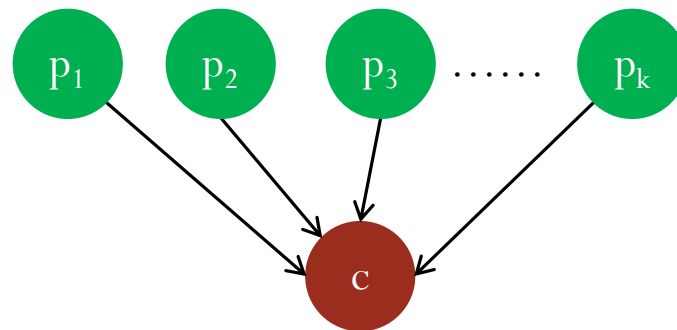


Outline

1. Introduction: TB and MTBC
2. Background: Post-genomic data analysis
3. TCF: Tensor Clustering Framework
4. Evolution model for spoligotypes
 - [Ozcaglar et al., IEEE BIBM 2011]
 - [Ozcaglar et al., IEEE Trans. NanoBioscience, to appear, 2012]
5. UBF: Unified Biclustering Framework
6. Conclusion

Motivation: Evolution of spoligotypes

- **Motivation:**
 - Putative mutation history of spoligotypes
 - Deletions in the DR region
 - Better understand the mutation mechanism of biomarkers
 - e.g. Rare convergent evolution in the DR region [Fenner et al, 2011]
- **Goal:** Disambiguate the ancestor spoligotypes

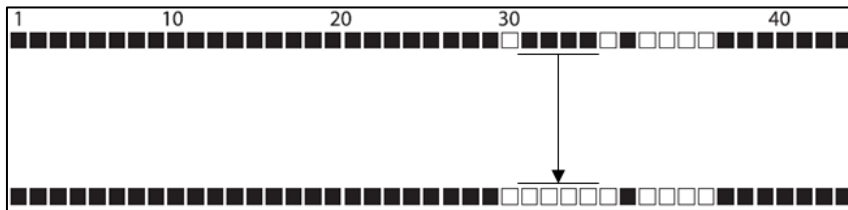


- **Method:** MakeSpoligoforest() algorithm
 - Uses an independent biomarker, MIRU-VNTR
 - Based on maximum parsimony

Mutation mechanism of biomarkers

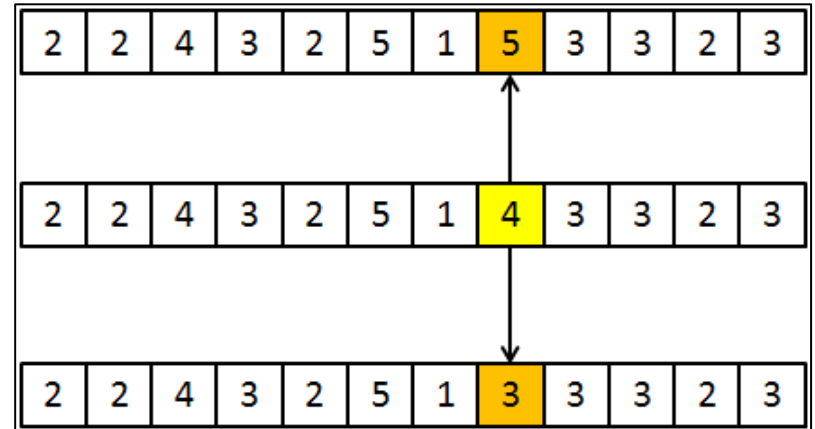
Spoligotype

- Unidirectional
- Spacers can be lost, but not gained
- Camin-Sokal parsimony
 - $1 \rightarrow 0$ ✓
 - $0 \rightarrow 1$ ✗
- Irreversible deletion
- *Contiguous deletion assumption (CDA)*



MIRU-VNTR

- Bidirectional
- Tandem repeats can be lost or gained
- Stepwise mutation model



Most parsimonious forest generation

- Assumptions
 - Contiguous deletion assumption
 - No convergent evolution
- Distance measures for strain comparison

- $\vec{s}_i \in \{0, 1\}$, where $i \in \{1, \dots, 43\}$
- $\vec{m}_j \in \{0, \dots, 15\} \cup \{s, t, \dots, z\}$, where $j \in \{1, \dots, 12\}$

1. Hamming distance between spoligotypes

$$H_S(\vec{s}_i, \vec{s}_j) = \sum_{r=1}^{43} |\vec{s}_{ir} - \vec{s}_{jr}|$$

2. Hamming distance between MIRU patterns

$$H_M(\vec{m}_i, \vec{m}_j) = \sum_{r=1}^{12} |\text{sign}(\vec{m}_{ir} - \vec{m}_{jr})|$$

3. L1 distance between MIRU patterns

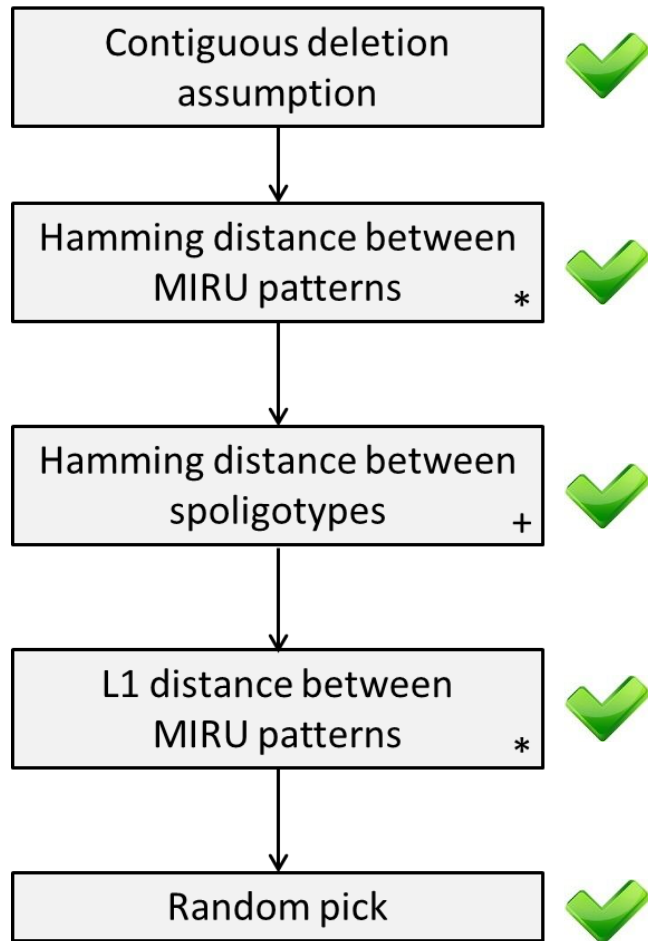
$$L_M(\vec{m}_i, \vec{m}_j) = \sum_{r=1}^{12} |\vec{m}_{ir} - \vec{m}_{jr}|$$

- Validation of the model

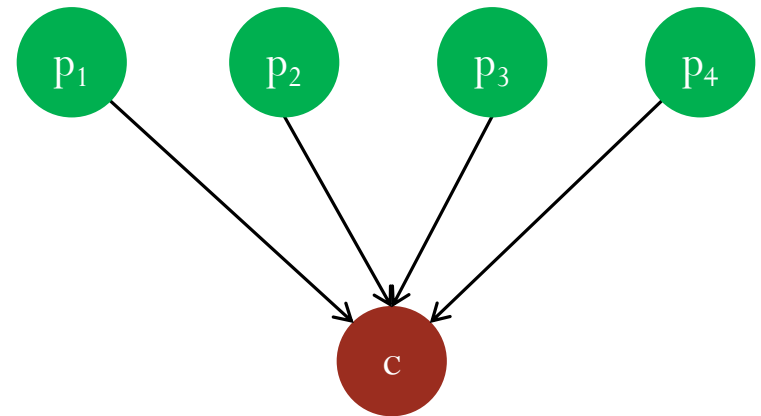
- Segregation accuracy: Percentage of within-lineage mutation events.

$$S = \frac{\sum_{l_i=l_j} d_{ij}}{\sum d_{ij}}$$

MakeSpoligoforest algorithm

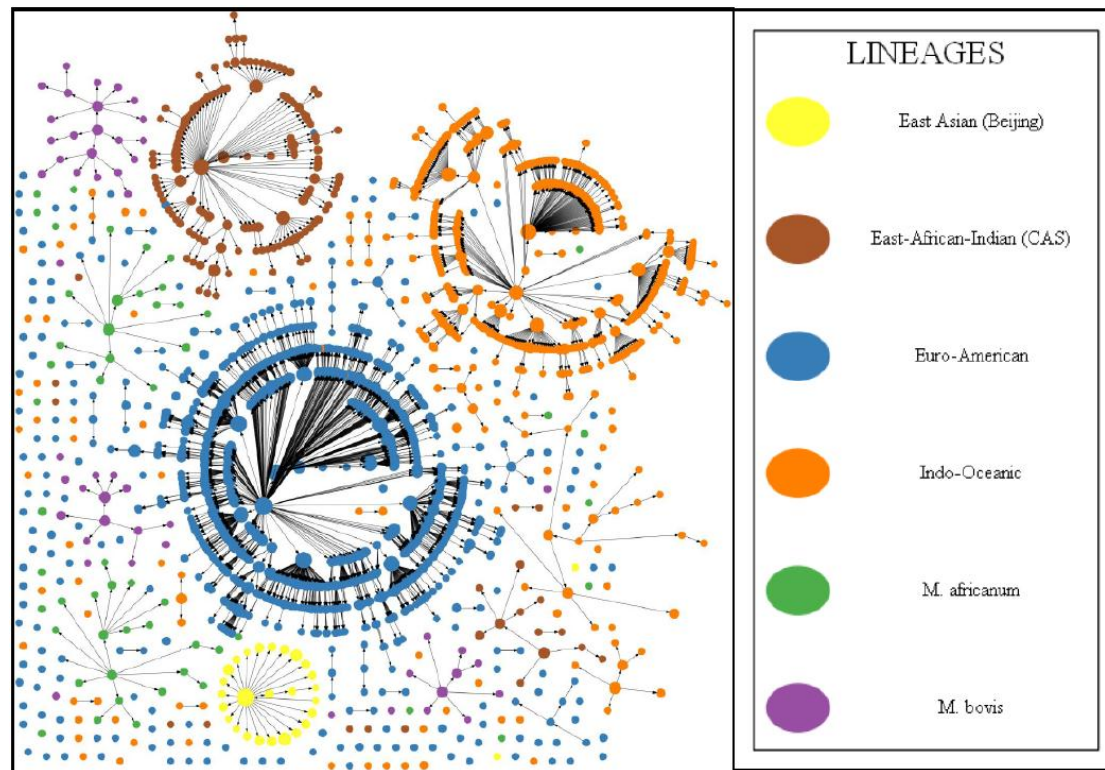


	P ₁	P ₂	P ₃	P ₄
H _M	1	2	1	1
H _S	3	2	5	3
L _M	6	3	4	6



The spoligoforest

- CDC dataset, 2004-2008
 - 9336 unique MTBC strains determined by spoligotypes and MIRU patterns
 - 2841 nodes: Spoligotypes
 - 2562 edges: Mutation events



Comparison with existing mutation models

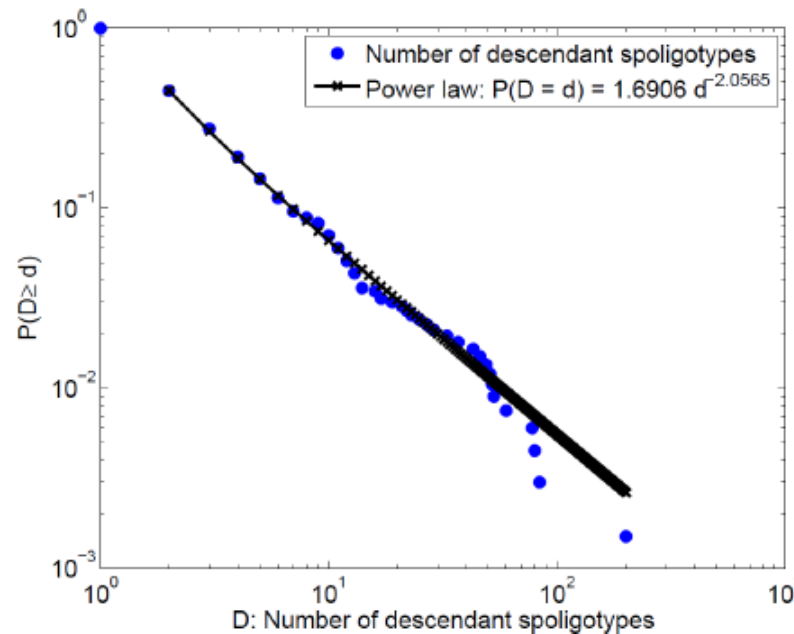
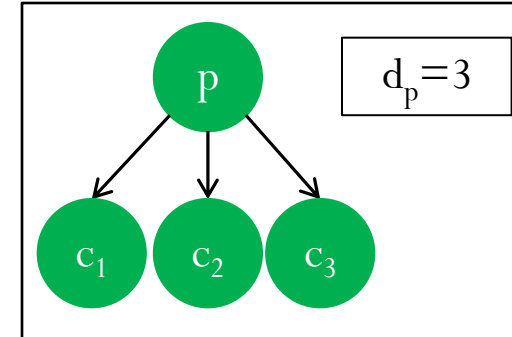
- The difference between segregation accuracy of different mutation models is not statistically significant

Model	Segregation accuracy	# Isolated nodes	# Mutation events
Zipf model [Reyes et al. 2008]	0.9921	235	2562
MakeSpoligoforest() (Spoligotype)	0.9906	230	2562
MakeSpoligoforest() (MIRU)	0.9941	233	2562
MakeSpoligoforest() (Spoligotype and MIRU)	0.9941	232	2562

- **MakeSpoligoforest() algorithm** results in similar percentage of within-lineage mutation events
- Alternative mutation models also perform as good
- We use the spoligoforest generated using **both biomarkers**

Result 1: Number of descendant spoligotypes

- d_i : Number of descendant spoligotypes of node i
- Number of descendant spoligotypes distribution
 - Power Law



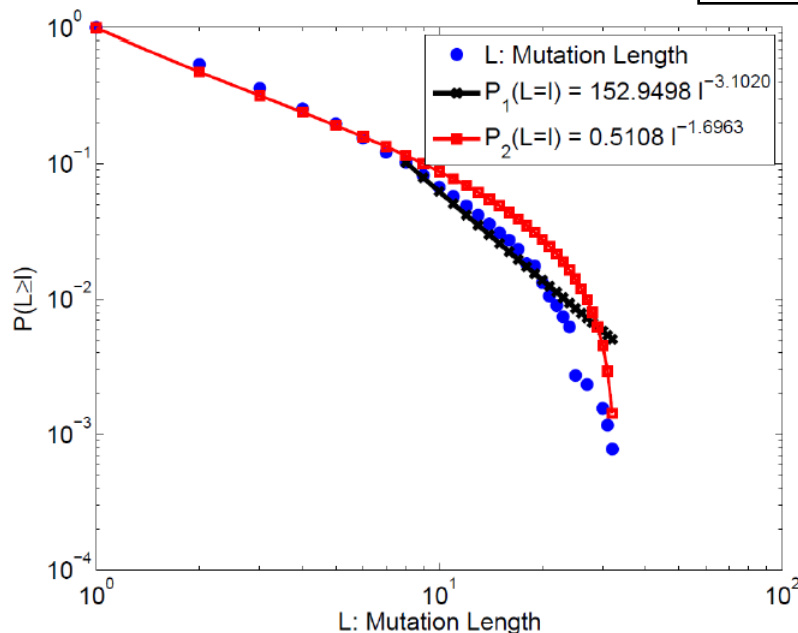
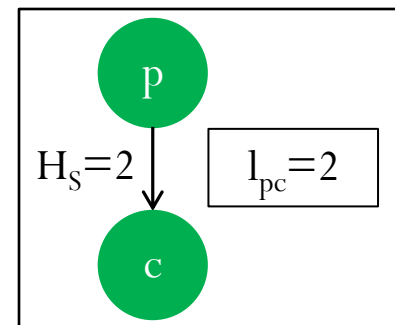
Clauset et al.,
SIAM Review, 2009.

Attribute	Power law distribution function	Domain	p -value	Support for power law
D: Number of descendant spoligotypes	$P(D = d) = 1.6906 d^{-2.0565}$	$d \geq 2$	0.6330	Good
L: Mutation length	$P_1(L = l) = 152.9498 l^{-3.1020}$	$l \geq 8$	0.0020	None
	$P_2(L = l) = 0.5108 l^{-1.6963}$	$1 \leq l \leq 43$	0	None

Result 2: Mutation length frequency

- **Mutation length:** Number of spacers deleted in a mutation
- l_{ij} : The length of mutation from node i to node j
- Zipf model by Reyes et al.

$$P_2(L = l) = \frac{l^{-\alpha}}{\sum_{i=1}^{43} i^{-\alpha}}$$

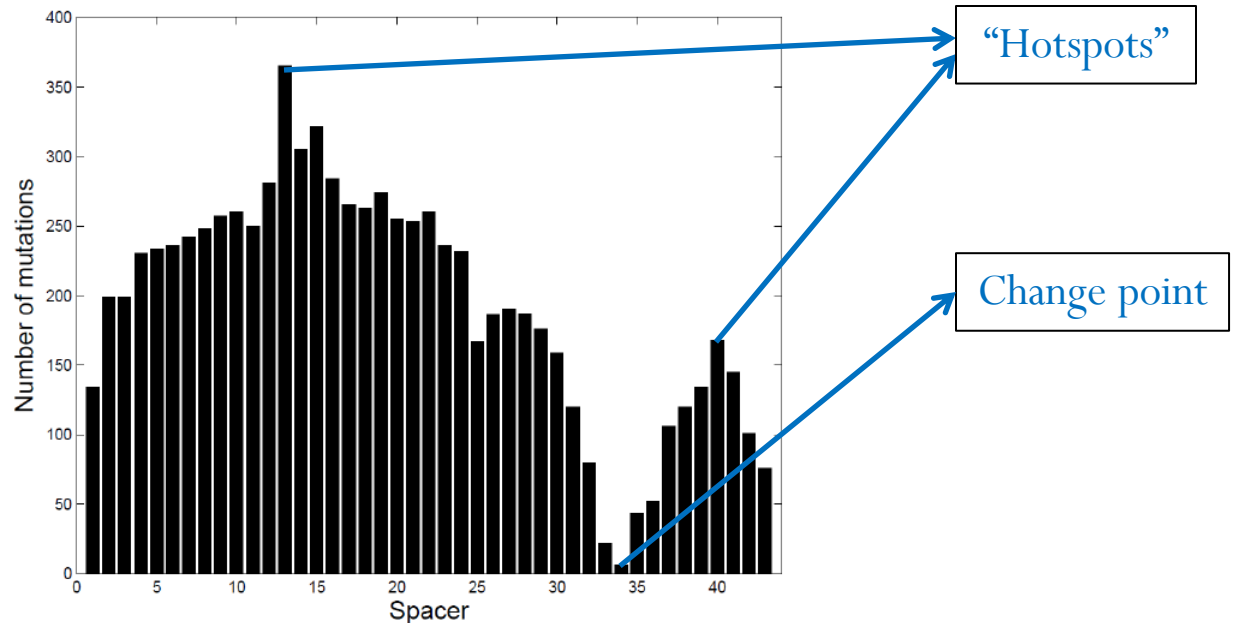


- Why not power law?
 - Longest observed mutation length: 32
 - Maximum possible mutation length: 43

Attribute	Power law distribution function	Domain	p-value	Support for power law
D: Number of descendant spologotypes	$P(D = d) = 1.6906 d^{-2.0565}$	$d \geq 2$	0.6330	Good
L: Mutation length	$P_1(L = l) = 152.9498 l^{-3.1020}$	$l \geq 8$	0.0020	None
	$P_2(L = l) = 0.5108 l^{-1.6963}$	$1 \leq l \leq 43$	0	None

Result 3: Number of mutations at each spacer

- Number of mutation events in which each spacer is deleted

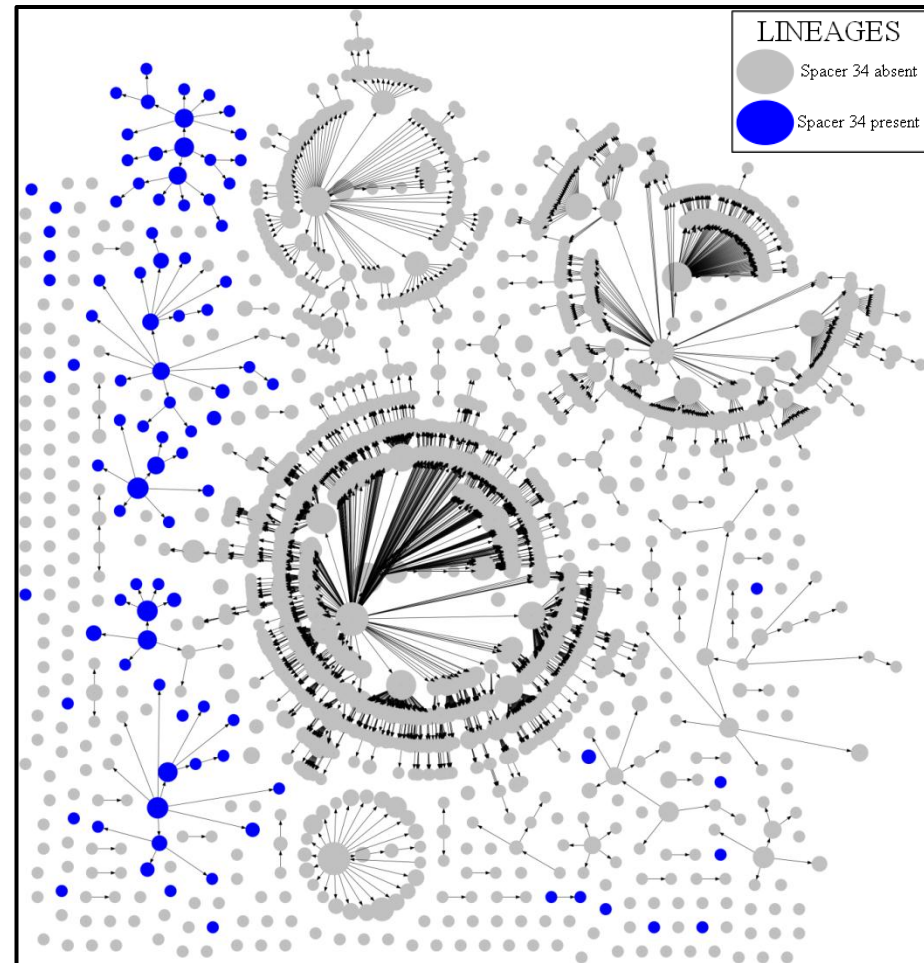


- Spatially bimodal distribution.
 - Hotspots, sites of increased observed variability: [Spacers 13 and 40](#).
 - Change point: [Spacer 34](#).

Ozcaglar et al., Inferred spoligoforest topology unravels spatially bimodal distribution of mutations in the DR region, *IEEE Trans. NanoBioscience*, in press, 2012.

Spatially bimodal distribution

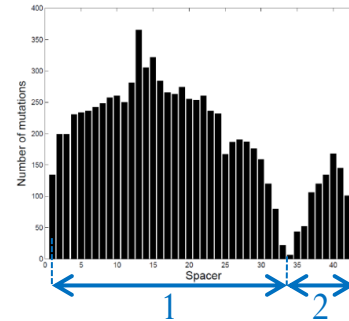
- Reason: Scarcity of sp33-sp36
- Proofs:
 - Principal genetic groups PGG 2 and PGG 3 defined by Sreevatsan et al. lack spacers 33 to 36.
 - Euro-American lineage is characterized by the deletion of spacers 33-36.
 - 1971 spoligotypes out of 2841, 69.37% in the CDC dataset are labeled with Euro-American lineage.
 - 94 out of 2841 spoligotypes, only 3.31% of them, have spacer 34 present in the DR region.



Result 4: Alternative model - SPM

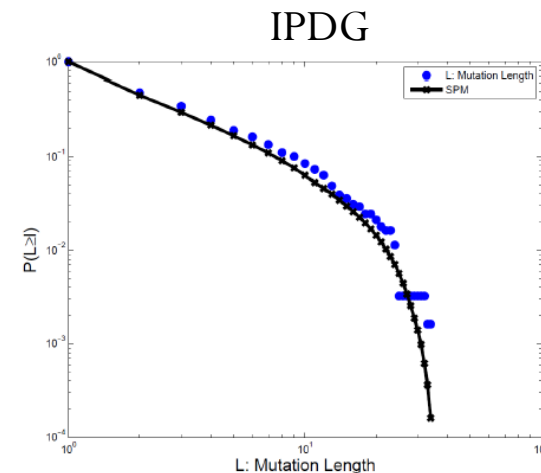
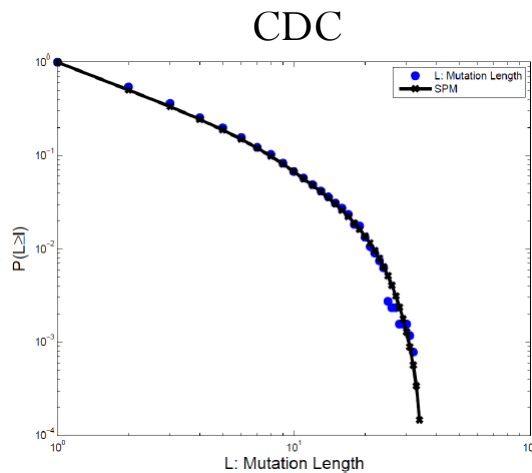
- SPM: Starting Point Model
 - Condition on the starting point of mutation

$$P(L = l | S = s) = \begin{cases} \frac{l^{-\alpha_s}}{\sum_{i=1}^{35-s} i^{-\alpha_s}}, & s \in [1, 33] \\ \frac{l^{-\alpha_s}}{\sum_{i=1}^{44-s} i^{-\alpha_s}}, & s \in [35, 42] \\ 1, & s \in \{34, 43\}, l = 1. \end{cases}$$



Start	End	Possible?
[1,34]	[1,34]	✓
[1,34]	[35,43]	✗
[35,43]	[1,34]	✗
[35,43]	[35,43]	✓

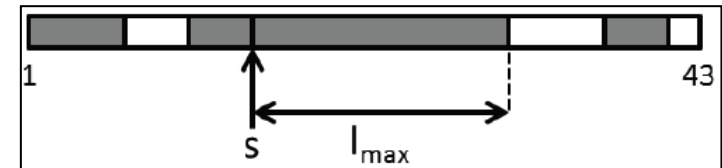
- SPM on mutation length frequency of CDC and IPDG datasets



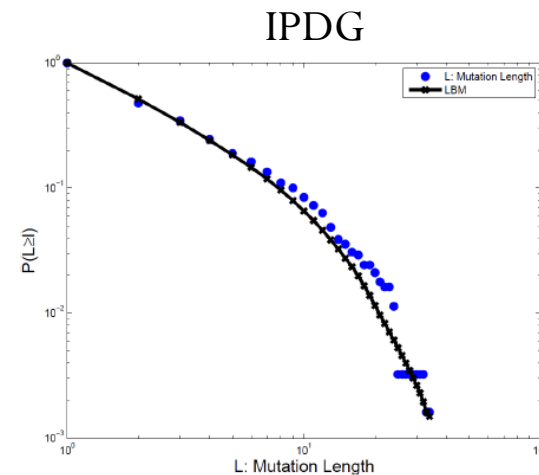
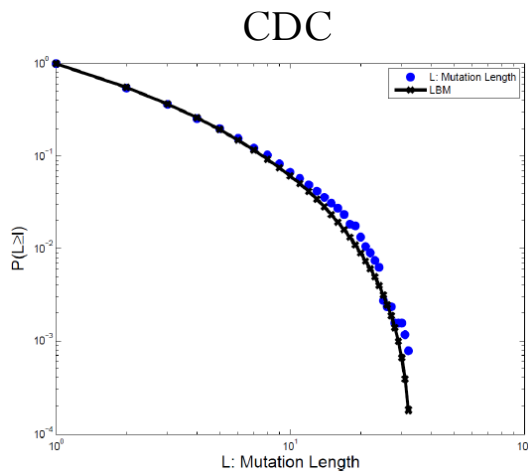
Result 4: Alternative model - LBM

- LBM: Longest Block Model
 - Condition on the length of longest block of spacers

$$P(L = l | L \leq l_{max}) = \begin{cases} \frac{l^{-\alpha l_{max}}}{\sum_{i=1}^{l_{max}} i^{-\alpha l_{max}}}, & l_{max} > 1, P(L \leq l_{max}) \neq 0 \\ 1, & l_{max} = l = 1. \end{cases}$$



- LBM on mutation length frequency of CDC and IPDG datasets



Outline

1. Introduction: TB and MTBC
2. Background: Post-genomic data analysis
3. TCF: Tensor Clustering Framework
4. Evolution model for spoligotypes
5. **UBF: Unified Biclustering Framework**
 - [Ozcaglar et al., RPI Technical Report, 2012]
6. Conclusion

Motivation and Goal: UBF

- Host-pathogen association analysis

- Stable: [Hirsh et al., *PNAS*, 2004]
- Variable: [Gagneux et.al., *PNAS*, 2006]

- Phylogeographic lineages:

- Genotype of MTBC and patient attributes are related

- MTBC strains: spoligotypes

- TB patients: country of birth

Genome-phenome data fusion

- Incorporate more data into domain knowledge

- Genetic distance between MTBC strains
- Spatial distance between TB patients
- Time of infection

Favor more likely mutation events

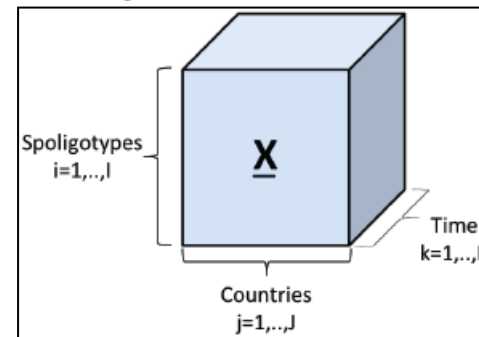
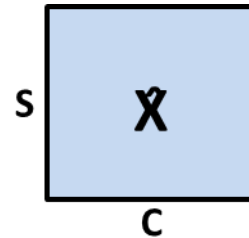
Favor more likely transmission events

Trace transmission routes

- **Need:** A framework to combine data from multiple sources

Biclustering problem

- Host-pathogen association analysis: a biclustering problem
- MTBC strains: spoligotypes
- TB patients: country of birth
- Dataset
 - NYC dataset: 4301 patients
 - 311 spoligotypes: KBBN, CBN
 - 104 countries
 - 7 years: 2001-2007
- Distance matrices



- Genetic proximity matrix

A screenshot of a software interface showing a genetic proximity matrix formula. The formula is:

$$P_G(s_i, s_j) = \begin{cases} \frac{1}{1 + H(s_i, s_j)}, & \text{if } i \neq j, \text{CDA}(s_i, s_j), H(s_i, s_j) \leq 10 \\ 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

The interface includes a "LINEAGES" tab and a "M. tou" button.

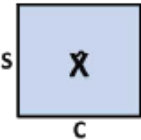
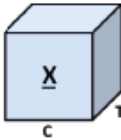
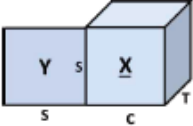
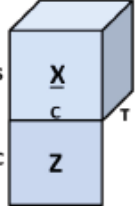
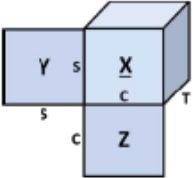
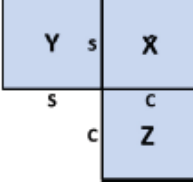
- Spatial proximity matrix

A screenshot of a software interface showing a spatial proximity matrix formula. The formula is:

$$P_S(C_i, C_j) = \begin{cases} \frac{1}{1 + L(C_i, C_j)}, & \text{if } i \neq j, L(C_i, C_j) \leq 3 \\ 1, & \text{if } i = j \\ 0, & \text{otherwise.} \end{cases}$$

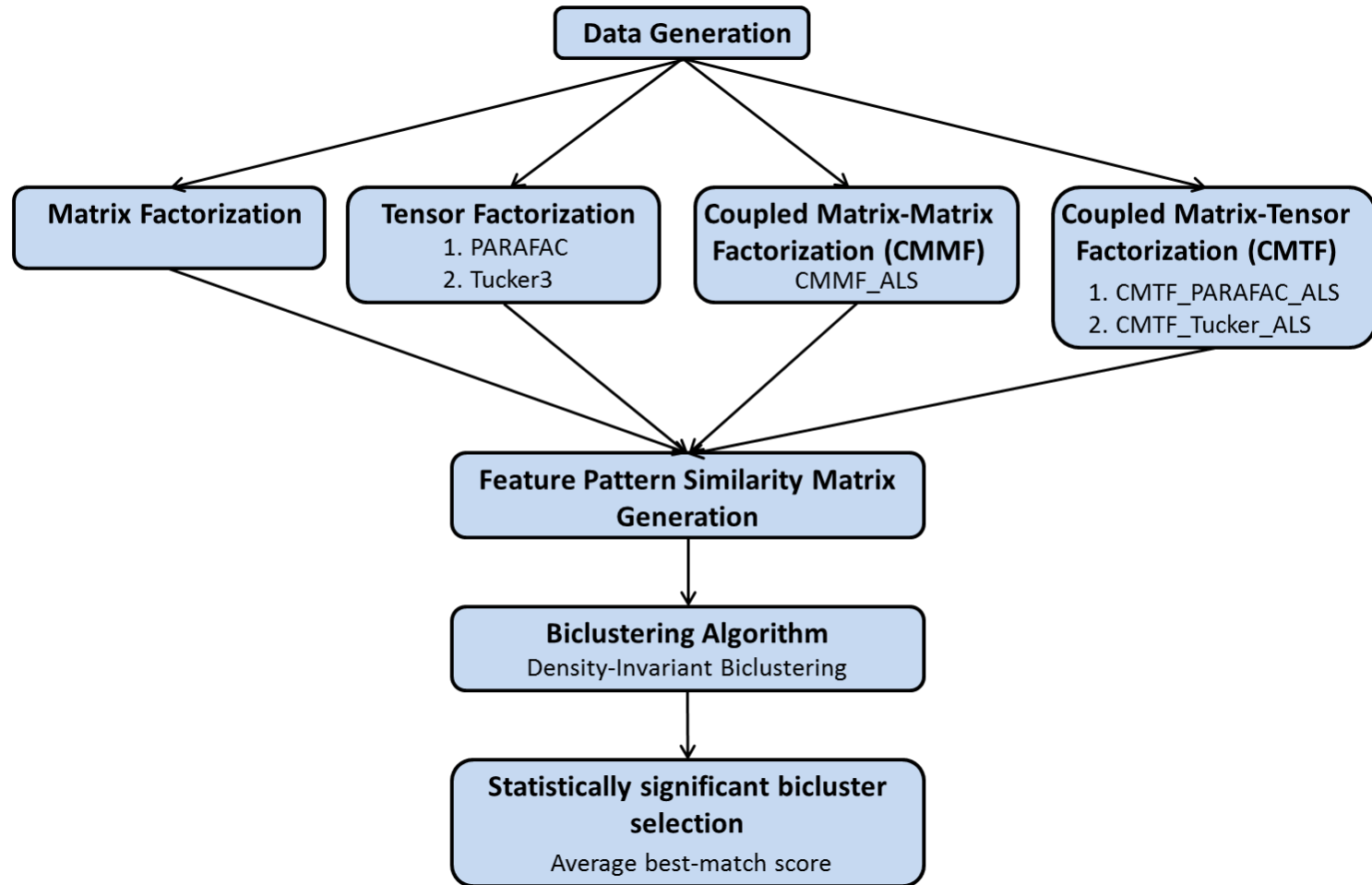
The interface includes a world map and a "M. tou" button.

Step 1: Data generation / fusion

Number	Data configuration	Extra information	Method in UBF
1		—	MBF
2		Time	TBF
3		Time + genetic distance	CMTBF _g
4		Time + spatial distance	CMTBF _s
5		Time + genetic distance + spatial distance	CMTBF _{gs}
6		Genetic distance + spatial distance	CMMBF

S: Spoligotype
C: Country
T: Time


UBF: Unified Biclustering Framework

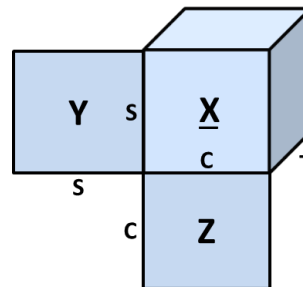
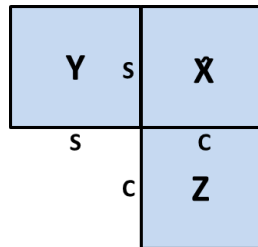
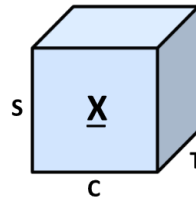
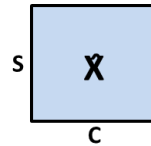


Ozcaglar et al., Host-pathogen association analysis of tuberculosis patients via Unified Biclustering Framework, *RPI Tech. Report*, 2012.

Step 2&3: Data factorization & FPSM generation

Data factorization

- Matrix factorization
 - The matrix itself
- Tensor factorization
 - PARAFAC
 - Tucker3
- Coupled matrix-matrix factorization
 - CMMF_ALS
- Coupled matrix-tensor factorization
 - CMTF_PARAFAC_ALS
 - CMTF_Tucker_ALS 



FPSM generation

- FPSM: Feature Pattern Similarity Matrix

- Calculation 1: Cosine similarity
 - PARAFAC, CMTF_PARAFAC_ALS
 - CMMF_ALS

$$\text{FPSM}_{ij} = \begin{cases} \frac{A_i \cdot B'_j}{\|A_i\| \|B_j\|}, & \text{if } N(i, j) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

- Calculation 2: Cosine similarity
 - Tucker3, CMTF_Tucker_ALS

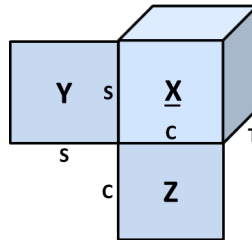
$$\hat{G}_{pq} = \sum_{r=1}^R G_{pqr}$$

$$\text{FPSM}_{ij} = \begin{cases} \frac{A_i \cdot \hat{G}}{\|A_i \cdot \hat{G}\|} \cdot \frac{B'_j}{\|B_j\|}, & \text{if } N(i, j) > 0 \\ 0, & \text{otherwise.} \end{cases}$$

CMTF_Tucker_ALS algorithm

- Algorithm 2 CMTF_Tucker_ALS($\underline{X} \in \mathbb{R}^{I \times J \times K}$, $\underline{Y} \in \mathbb{R}^{I \times M}$, $\underline{Z} \in \mathbb{R}^{J \times N}$, $[P, Q, R]$)

- $$\begin{aligned} X_{(1)} &\approx A G_{(1)} (C' \otimes B') \\ Y &\approx A V' \\ Z &\approx B W' \end{aligned}$$



- $$L_3 = \|X_{(1)} - A G_{(1)} (C' \otimes B')\|_F^2 + \|Y - A V'\|_F^2 + \|Z - B W'\|_F^2$$

- $$\begin{aligned} \min_A & -\text{tr}(A' M M' A) - \text{tr}(A' Y Y' A) \\ \text{s.t. } & A' A = I \end{aligned} \quad \Rightarrow \quad A = \text{EVD}(M M' + Y Y', P)$$

$$M = X_{(1)} (C C' \otimes B B')$$

- $$\begin{aligned} \min_B & -\text{tr}(B' M_2 M_2' B) - \text{tr}(B' Z Z' B) \\ \text{s.t. } & B' B = I \end{aligned} \quad \Rightarrow \quad B = \text{EVD}(M_2 M_2' + Z Z', Q)$$

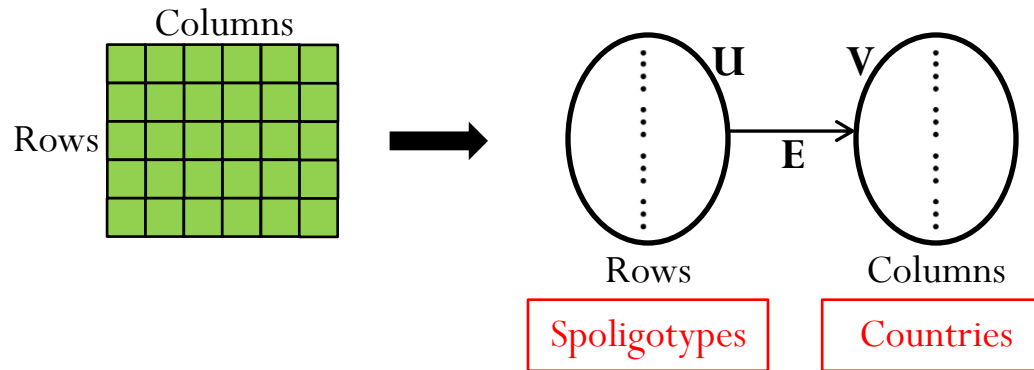
$$M_2 = X_{(2)} (C C' \otimes A A')$$

- $$\begin{aligned} \min_C & -\text{tr}(C' M_3 M_3' C) \\ \text{s.t. } & C' C = I \end{aligned} \quad \Rightarrow \quad C = \text{SVD}(M_3, R)$$

$$M_3 = X_{(3)} (B B' \otimes A A')$$

Step 4: Density-invariant bicluster

- **Bicluster** $B = (U, V, E)$ as a **bipartite graph** $G = (U, V, E)$



- **Density** and **variance** of a graph

$$d(G) = \frac{\sum_{e \in E} w(e)}{\binom{|V|}{2}}$$

$$v(G) = \sqrt{\frac{1}{|E| - 1} \sum_{e \in E} (w(e) - \bar{w})^2}$$

- **Density-invariant bicluster**

$$\begin{aligned} &1. d(B) \geq \alpha, v(B) \leq \beta \\ &2. d(B') \geq \alpha, v(B') \leq \beta \quad \forall B' = B \setminus \{m\} \text{ where } m \in U \cup V, |B'| > 0 \end{aligned}$$

Step 4&5: Density-invariant biclustering

- **Density-invariant biclustering algorithm (DIB)**

1. Discretize \mathbf{X} with threshold th

$$D_{ij} = \begin{cases} 1, & \text{if } X_{ij} \geq th \\ 0, & \text{otherwise.} \end{cases}$$

2. Find candidate biclusters using BiMax [Prelic et al, 2006]

$$\text{CandidateBiclusters} = \text{BiMax}(\mathbf{D})$$

3. Find (α, β) -density-invariant biclusters among candidate biclusters

- **Statistically significant bicluster selection**

- For two biclusters $B_1 = (G_1, C_1)$ and $B_2 = (G_2, C_2)$

$$\text{match}(B_1, B_2) = \frac{|G_1 \cap G_2| + |C_1 \cap C_2|}{|G_1 \cup G_2| + |C_1 \cup C_2|}$$

- **Stable bicluster:** Average best-match stability ≥ 0.95

Results – Biclusters based on KBBN sublineages

Bicluster	Number of patients	Spoligotypes			Countries	
		STT no	Major lineage	Sublineage	Name	TB continent
B14	6	ST1908 ST58	Euro-American Euro-American	H3 T5	Ecuador	Americas
B16	2	ST897	Indo-Oceanic	EAI2-Manila	Philippines	Southeast Asia
B321	32	ST55 ST62 ST51 ST1908	Euro-American Euro-American Euro-American Euro-American	T1 H1 T1 H3	Ecuador	Americas
B421	32	ST1	East Asian	Beijing	Taiwan Barbados Dominica Malaysia Myanmar Philippines	East Asian Americas Americas Southeast Asia Southeast Asia Southeast Asia
B422	27	ST1 ST38	East Asian Euro-American	Beijing X2	Malaysia Philippines	Southeast Asia Americas
B525	11	ST167 ST42 ST57 ST904 ST904 ST904	Euro-American Euro-American Euro-American Euro-American <i>M. africanum</i> <i>M. africanum</i>	T1 LAM9 LAM10-CAM T5 AFRL1 AFRL1	Haiti	Americas
B64	3	ST1391 ST58	Indo-Oceanic Euro-American	EAI5 T1	Bangladesh	Indian Subcontinent

1. Philippines: EAI2_Manila strain ST897
2. East Asian Beijing strain ST1: three TB continents. Transmissible.
3. Malaysia & Philippines: ST1 and ST38. Neighbour countries.



Results – Biclusters within each CBN lineage

Bicluster	Number of patients	Spoligotypes			Countries	
		SIT no	Major lineage	Sublineage	Name	TB continent
B712	5	UST251 ST478 ST1154	Euro-American Euro-American Euro-American	S X2 LAM9	Mexico	Americas
B732	9	ST471 ST25 ST381 ST21 ST203 UST167	East-African Indian East-African Indian East-African Indian East-African Indian East-African Indian East-African Indian	CAS1-Delhi CAS1-Delhi CAS1-Delhi CAS CAS EAI5	China	East Asia
B733	11	ST381 ST25 ST21 UST167	East-African Indian East-African Indian East-African Indian East-African Indian	CAS1-Delhi CAS1-Delhi CAS EAI5	China Dominican Republic	East Asia Americas
B741	7	ST1162 ST941 ST541 ST1168	East Asian East Asian East Asian East Asian	Beijing Beijing Beijing Beijing	Haiti	Americas
B742	212	UST1 ST255 ST260 ST941 ST265 ST190 ST1	East Asian East Asian East Asian East Asian East Asian East Asian East Asian	Beijing Beijing Beijing Beijing Beijing Beijing Beijing	United States	Americas
B743	291	ST200 ST265 ST1	East Asian East Asian East Asian	Beijing Beijing Beijing	China United States	East Asia Americas
B751	17	ST325 ST326 ST187 ST181 ST319 ST331 UST229	<i>M. africanum</i> <i>M. africanum</i> <i>M. africanum</i> <i>M. africanum</i> <i>M. africanum</i> <i>M. africanum</i> <i>M. africanum</i>	AFRL1 AFRL1 AFRL1 AFRL1 AFRL2 AFRL2 AFRL2	United States	Americas
B761	3	ST479 ST481	<i>M. bovis</i> <i>M. bovis</i>	BOV BOV_1	Dominican Republic	Americas
B762	9	ST409 ST683	<i>M. bovis</i> <i>M. bovis</i>	BOV_2 BOV_2	United States	Americas

Outline

1. Introduction: TB and MTBC
2. Background: Post-genomic data analysis
3. TCF: Tensor Clustering Framework
4. Evolution model for spoligotypes
5. UBF: Unified Biclustering Framework
6. **Conclusion**

Conclusion

1. TCF: Tensor Clustering Framework

- Genomic data fusion via MBT: multiple-biomarker tensor
- Simultaneous analysis of two biomarkers
- A new sublineage structure of MTBC based on multiple biomarkers
- Divided, merged, or validated existing sublineages

2. Evolution of spoligotypes

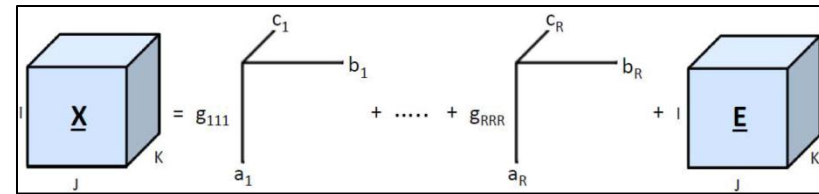
- Genomic mutation mechanism fusion via an additional biomarker
- Number of descendant spoligotypes follows power law
- Number of mutations at each spacer follows a spatially bimodal distribution
- Mutation length frequency does not follow power law. Alternatives:
 - SPM: Starting Point Model
 - LBM: Longest Block Model

3. UBF: Unified Biclustering Framework

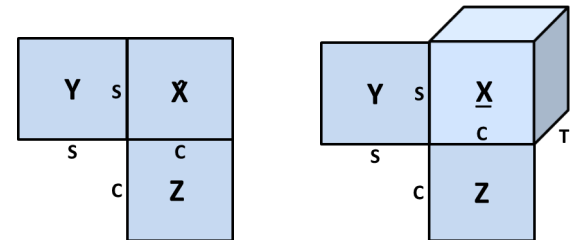
- Genome-phenome data fusion
- Incorporate genetic distance, spatial distance and time
- Found existing and new host-pathogen associations

Future work

- Non-deterministic tensor decomposition
 - Initial algorithm: Simulated Annealing with Adaptive stepsize (SAAS)
 - Tensors with varying size, rank, collinearity, noise level
 - Challenges: Global minima, overfactoring
 - Model selection framework for different types of noise
 - New constraints: sparsity, non-negativity



- Host-pathogen association analysis
 - Additional MTBC biomarkers: MIRU-VNTR, RFLP
 - Additional patient attributes: age group, homelessness, HIV status
 - Immigration map instead of world map
 - Line-search for ALS-based coupled factorization algorithms
 - Faster convergence to more accurate solutions



Acknowledgements

- My advisor
 - Prof. Bulent Yener
- Committee members
 - Prof. Kristin Bennett
 - Prof. Mohammed Zaki
 - Prof. Chris Bystroff
 - Prof. Qiang Ji
- Colleagues
 - Amina Shabbeer
 - Dr. Minoo Aminian
- This work was made possible by CDC and NIH



Publications used in this thesis

- Survey

- C. Ozcaglar, A. Shabbeer, S. L. Vandenberg, B. Yener, and K. P. Bennett, “**Epidemiological models of *Mycobacterium tuberculosis* complex infections**”, *Mathematical Biosciences*, vol. 236, no. 2, pp. 77-96, 2012. **Most accessed paper of Mathematical Biosciences journal in March-June 2012.**

- TCF

- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, and K. P. Bennett, “**Sublineage structure analysis of *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors**,” *BMC Genomics*, vol. 12, no. Suppl 2, p. S1, 2011.
- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, and K. Bennett, “**Examining the sublineage structure of *Mycobacterium tuberculosis* complex strains with multiple-biomarker tensors**,” in 2010 *IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, pp. 543-548, 2010.
- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, K. P. Bennett, “**Multiple-biomarker tensor analysis for tuberculosis lineage identification**,” *NIPS Workshop on Tensors, Kernels and Machine Learning*, 2010.
- C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, K. P. Bennett, “**A clustering framework for *Mycobacterium tuberculosis* complex strains using multiple-biomarker tensors**”, Rensselaer Polytechnic Institute. TR-10-08, 2010.

Publications used in this thesis & Software

- Evolution model of spoligotypes

- C. Ozcaglar, A. Shabbeer, N. Kurepina, N. Rastogi, B. Yener, and K. P. Bennett, “**Inferred spoligoforest topology unravels spatially bimodal distribution of mutations in the DR region**,” *IEEE Trans. NanoBioscience*, 2012.
- C. Ozcaglar, A. Shabbeer, N. Kurepina, B. Yener, and K. Bennett, “**Data-driven insights into deletions of *Mycobacterium tuberculosis* complex chromosomal DR region using spoligoforests**,” in 2011 *IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM)*, pp. 75-82, 2011.

- UBF

- C. Ozcaglar, B. Yener, and K. P. Bennett, “**Host-pathogen association analysis of tuberculosis patients via unified biclustering framework**,” Tech. Rep. 12-05, Department of Computer Science, Rensselaer Polytechnic Institute, 2012.

- Software

- TCF
- Spoligoforest generator
- UBF

Publications not used in this thesis

- A. Shabbeer, C. Ozcaglar, B. Yener, K. P. Bennett. **Web tools for molecular epidemiology of tuberculosis.** *Infection, Genetics and Evolution*, 2011. **Most accessed paper of Infection, Genetics and Evolution journal as of December 2011.**
- K. P. Bennett, C. Ozcaglar, J. Ranganathan, S. Raghavan, J. Katz, D. Croft, B. Yener, A. Shabbeer. **Visualization of tuberculosis patient and *Mycobacterium tuberculosis* complex genotype data via host-pathogen maps.** *IEEE BIBM Workshop on Computational Advances in Molecular Epidemiology*, 2011.
- M. Aminian, A. Shabbeer, K. Hadley, C. Ozcaglar, S. Vandenberg, K. P. Bennett. **Knowledge-based Bayesian network for the classification of *Mycobacterium tuberculosis* complex sublineages.** *ACM BCB*, 2011.
- M. Aminian, A. Shabbeer, K. Hadley, C. Ozcaglar, S. Vandenberg, K. P. Bennett. **Incorporating biology rules of thumb into Bayesian networks.** *J. Computational Biology and Bioinformatics*, in press, 2012.
- A. Shabbeer, C. Ozcaglar, M. Gonzalez, K. P. Bennett, **Optimal Embedding of Heterogeneous Graph Data with Edge Crossing Constraints.** *NIPS Workshop on Challenges of Data Visualization*, 2010.
- A. Shabbeer, L. S. Cowan, C. Ozcaglar, N. Rastogi, S. L. Vandenberg, B. Yener, and K. P. Bennett, "TB-Lineage: An online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex," *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 789-797, 2012.
- A. Shabbeer, C. Ozcaglar, K. P. Bennett, **Crossing minimization within graph embeddings.** *Submitted to Journal of Machine Learning Research*.
- J. M. Pyle, F. S. Spear, S. Adali, B. Szymanski, S. Pearce, A. Waters, Z. Linder, C. Ozcaglar, **MetPetDB: The unique aspects of metamorphic geochemical data and their influence on data model, user interface and collaborations.** *Geological Society of America Abstracts with Programs*, 2007.

Thank you